# THE SERVICE OF STATISTICS TO SOCIOLOGY*

By Franklin H. Giddings, Ph.D., LL.D.,
*Professor of Sociology, Columbia University.*

Statistics have been of relatively limited service to sociology hitherto because of the quantitative limitations of sociological data.

In economics we have not only frequencies of Sort or Kind, but also abundant frequencies of Size. Prices of commodities, wage rates, interest rates, rents, dividends, and tax rates are measures of size as well as categories of kind. In sociology, while frequencies of size are by no means wanting, as witness, birth and death rates, marriage and divorce rates, by far the greater proportion of our numerical data are frequencies of sort or kind. For example, respective numbers of the different nationalities entering the United States through our ports of immigration, numbers of the foreign and the native born, numbers of the literate and the illiterate, numbers of adherents to the different religious creeds, numbers of persons engaged in various occupations, numbers of the delinquent and the dependent, and so on are frequencies of sort.

Practically all of our statistical operations have been developed through comparisons of size and analyses of size frequencies. Normal frequency distribution, mean square deviation, the coefficient of variability, probable error, and the coefficient of correlation are measures inherent in items of size rather than in items of sort.

Perhaps it has been too hastily assumed that statistical results obtainable through the use of these measures can never be obtained in any other way; and it may be worth while to indicate certain possibilities of statistical measurement that lie in frequencies of sort which, if systematically applied to our large collections of data in census and other reports, might add much to our scientific knowledge of social relations.

Let us then for a moment examine certain statistical measures that may be derived from the mere inequality of sort frequencies; such inequality, for example, as is presented by the num-

* Paper read at the seventy-fifth anniversary meeting of the American Statistical Association, Boston, Mass., February 14, 1914.

bers of foreign-born Irish, foreign-born Germans, foreign-born French Canadians, foreign-born Italians, and so on, in the population of any one of our commonwealths.

The inequality of two numbers a and b is measured by the difference a−b. The total inequality of n numbers, one to another, is the sum of the differences found when each number is subtracted from every other number, and each difference is counted once, or:

Total inequality of n numbers = (a−b)+(a−c) + (a−d) +.....+(a−n) + (b−c) + (b−d) +..... + (b−n) + (c−d) + ..... + (c−n) + ..... + (d−n).

If the smallest of n numbers be subtracted from the largest, the difference, m−s is the *range* of inequality. It is the quantity of reference from which a measure of tendency to inequality among given numbers—or frequencies—may be derived.

If each lesser frequency in turn be subtracted from a maximum frequency the differences will all be positive. Their sum is the positive inequality of the frequencies in relation to the largest frequency among them, and this, divided by the number of frequencies, is the mean positive inequality.

If each greater frequency in turn be subtracted from a minimum frequency the differences will all be negative. Their sum is the negative inequality of the frequencies in relation to the smallest frequency among them, and this, divided by the number of frequencies, is the mean negative inequality.

If the ascending or descending steps of arrayed frequencies are equal, the positive and negative inequalities, as above defined, are equal. The mean inequality then is equal to one half of the difference between the maximum and the minimum frequency, or to one half of the range of inequality.

If the steps of arrayed frequencies are unequal, the positive inequality will be greater or less than the negative inequality, but the mean of the arithmetical values of the positive and negative averages will equal as before one half of the range of inequality.

The mean inequality of frequencies, of which the positive and the negative inequalities are equal, may be called *iota*, $\iota$. The mean positive inequality of frequencies may then be expressed as $\iota_1$, and the mean negative inequality may be expressed as $\iota_2$.

If the negative inequality of frequencies is equal to their positive inequality, the strength of tendency to equality among them may be described as equal to the strength of tendency among them towards inequality.

If the negative and the positive inequalities of frequencies are unequal, $\iota - \iota_1$ equals $\iota - \iota_2$, but the signs of these differences are opposite. If $\iota - \iota_1$ is positive, it measures strength of tendency towards equality; if negative, it measures strength of tendency towards inequality. If, $\iota - \iota_2$ is negative, it measures strength of tendency towards equality; if positive, it measures strength of tendency towards inequality.

If the negative and the positive inequalities of frequencies are equal $\frac{\iota}{\Sigma f}$ is a coefficient of inequality; if they are approximately equal it is an approximate coefficient of inequality.

In any case, $\frac{\iota - \iota_1}{\Sigma f}$, or $\frac{\iota - \iota_2}{\Sigma f}$ is a precise coefficient of strength of tendency towards equality or towards inequality, according to sign.

These relations stand out sharply in graphic presentation.

If the positive and the negative inequalities of frequencies are equal, the frequencies, when plotted as equi-distant ordinates arrayed, will make a figure that may be bounded at the top by a straight slant line.

If the positive inequality of frequencies exceeds the negative inequality, the frequencies, when plotted as equi-distant ordinates arrayed, will make a figure that may be bounded at the top by a downward curving or concave line.

If the negative inequality of frequencies exceeds the positive inequality the frequencies when plotted as equi-distant ordinates arrayed, will make a figure that may be bounded by an upward curving or convex line.

From mere frequencies of sort we can obtain also a measure of the sociologically important phenomenon of coördination, as distinguished from correlation.

Coördination is equivalence of position. For example, in botanical or in zoölogical classification genera are coördinate one with another, but are subordinate to orders as orders are to classes; species are coördinate one with another, but are subordinate to genera. In the ecclesiastical hierarchy priests are

of coördinate rank, bishops of coördinate higher rank, and archbishops of coördinate rank yet higher.

Superordinated or subordinated coördination, or the coördination of units within each rank throughout a succession of ranks, one above or one below another, is obviously a phenomenon incidental to all subclassification, creating intra-secting categories, or category within category, in descending comprehensiveness.

And when categories are intra-secting, the whole content of category B falls within category A; the whole content of category C falls within category B; and so on.

Therefore, statistically, coördination, superordinate or subordinate, is the appearance of certain same units in each of n categories.

Identity or sameness of content in each of two or more categories may be called Categorical Solidarity.

Since all units of category B, intra-secting category A, occur also in A, it is plain that in these two categories taken together there are as many unit instances of "same content" in more than one category as there are units in category B.

All units of category C, intra-secting B, occur also in B and in A. Therefore, taking the first three categories together, they present as many unit instances of "same content" in more than one category as there are units in B plus the number of units in C.

All units of category D occur also in C, in B, and in A. Therefore, in these four categories taken together there are as many unit instances of "same content" in more than one category as there are units in B, plus units in C, plus units in D.

In general, if categorical solidarity be expressed by S, a comprehensive category by K, and intra-secting categories by $k_1$, $k_2$, $k_3$, . . . . . , $k_n$

$$S = k_1 + k_2 + k_3 + \ldots . + k_n$$

S, so obtained, is an amount, and it is affected by the number of categories used. The degree, or average density, of solidarity, category with category, may be obtained, therefore by dividing S by n.

If the number of units in each of n categories were the same, and if the "same content" (neither more nor less, nor different) were in each and all categories, K would equal $k_1$, would equal

$k_2$, would equal $k_3$, and so on. Indicating the greatest arithmetically possible solidarity of n categories by G, we of course have $G = K(n-1)$, and the greatest possible degree of solidarity of n categories is $\frac{G}{n}$.

The ratio $\frac{S}{n} : \frac{G}{n}$ or $\frac{S}{G}$ is the coefficient of solidarity for any values of K, $k_1$ and n.

All measures of solidarity are corresponding measures of coördination.

With these additions to our means of measurement, what are the possibilities of statistical analysis in sociology?

Any association of units presents to the observer certain aspects which admit of quantitative description by statistical methods.

These aspects are: 1, Extent; 2, Duration; 3, Strength; 4, Compositeness; 5, Form; 6, Reaction; 7, Central Point or "center of gravity" of Reaction; 8, Contingency.

The statistical examination of the extent and the duration of association is the simplest of statistical operations. It involves only completeness and accuracy of count, and accurate determinations of date.

Strength of association is resistance to dissolution or disintegration. Dissolution, or disintegration, is statistically measured by the percentage, or other proportion, of associations of a given kind that break up within a given time. Family cohesion, for example, is measured by the divorce rate.

When units of more than one sort are combined in a mixture, the compositeness thereby arising is of three degrees, which may be named respectively, Variegation, Approximate Composition, and True Composition or Heterogeneity.

Variegation is determined by two variable quantities only, namely, (1) the number of sorts in the composition, and (2) the number of items in each sort, that is, the frequencies of the sorts. Differences of magnitude among variants (*i. e.*, units of sort) and the amount of difference that exists between any one sort and any other sort (*i. e.*, inequalities of interval or step) are neglected.

When the categorical or sort frequencies of a composition

are approximately equal, the variegation may be described as uniform.

When sort frequencies are unequal, and one frequency exceeds any other, the variegation thence resulting may be described as modal.

Variegation is measured, and thereby quantitatively described, by the coefficient of tendency towards equality, or towards inequality, as the case may be, of the sort frequencies, of the composition.

Approximate composition takes account of the difference between each frequency and every other frequency in the composition. It is measured by the total inequality of the frequencies.

True composition, or heterogeneity, is the totality of differences in a composition. It includes not only frequencies of sort, and inequalities of frequencies one to another, but also all differences of item from item (in respect of dimension, weight, value, or other magnitude), and all differences of interval or step. If the data are known, heterogeneity can be computed by simple algebraic methods, which are, however, tedious.

Variegation, fortunately the simplest phase of compositeness, is a fact of significance for the organic and social sciences. Easily measured, it is a measure itself, of strength of tendency, or of influences selective or constraining.

Average deviation and standard deviation are assumed to measure the strength of a mode-making tendency, selection, or pressure acting upon variates, *i. e.*, units of size.

If, for example, poppy capsules be gathered at random from a field, and the number of stigmatic bands on each capsule be counted, and the deviation from the mean number be found to be very small, the fact is supposed to tell us that the poppies in that field, or their progenitors, have survived a severe natural selection. The smaller the standard deviation, it is inferred, the greater has been the selective or mode-making pressure.*

The same significance attaches to variegation. The coefficient of tendency to or from equality is a measure of mode-making tendency, selection or pressure for frequencies of sort, probably quite as trustworthy as the coefficient of variability for frequencies of size.

* *Vidi* Karl Pearson, "Grammar of Science," Second Edition, Chapter X, sec. 5.

If seeds of a dozen kinds be planted simultaneously and indiscriminately, but in equal numbers, in a patch of garden, which is then neglected, and six weeks later plants of the dozen kinds are flourishing in approximately equal numbers, kind for kind, we infer that no selective influence has affected them; while if plants of one or two kinds at the end of the six weeks are relatively numerous, of other kinds relatively few, and of the remaining kinds very few, we infer that selection has been rigorous.

Uniformity of variegation then, means a negligible mode-making tendency, selection, or pressure; while marked modality of sort frequencies means a mode-making tendency relatively strong, or a mode-making selection or pressure relatively severe.

The Forms of association are (1) Tangent, or exclusive, no unit of one association occurring in another association; (2) Inter-secting, certain units occurring in more than one association but no association being wholly comprised in another; and (3) Intra-secting, all the units of association B occurring in association A, all the units of association C occurring in association B, and so on in descending comprehensiveness.

Tangent association is otherwise described as "segregation" when the units of each association are similar. The simple statistical problem presented is to count the number of like units that in one or another way are placed or combined in exclusive grouping.

Intra-secting associations are a case of subclassification and coördination, superordinate or subordinate. The coefficient of coördination is a measure of intra-secting association.

The Reactions of Association are measured in units of time, of displacement, and of transformation. Promptness and persistence of reaction are measured in units of time. Degree, extent and amount of reaction are measured in units of displacement or transformation. The statistical description of these reactions involves no unusual developments of method. The difficulties that are encountered arise in the determination of data, in making the original measurements.

The Central Point of Reaction, or, using a figure of speech, the center of gravity of reaction, is that point about which all reactions, including opposing ones, are in equilibrium. If units

react in different ways and with equal power or "weight" the center of reaction is the median of the array of the units. If the several units, either individually or when massed in those squadrons of units which we call frequencies, react with different power or "weight," the center of reaction is found on that side of the median where the heavier weighting occurs. The statistical problems, accordingly, that arise in any attempt to determine the center of gravity of associational reaction are those which in statistical analysis are known as questions of weighting.

No phenomena of society are of greater interest than are the shiftings of the centers of associational reaction. Among these are the shifting of the center of gravity in politics between opposing parties, between radicals and conservatives, between classes and masses, between rationalists and the upholders of instituted authority.

The foregoing aspects of association are of interest in themselves and also, in a higher degree, because of their relation to Contingency. In determining how far association or any phase of it is quantitatively linked with any other fact, we get close to the problems of law and cause. Contingency is measured by a percentage or by the Pearsonian or other coefficient of contingency. Actual contingency when found should be compared with a theoretically probable contingency.

The contingency of any phenomenon of association may be with an extraneous fact, or with any other phenomenon of association itself. Extraneous facts collectively are the environment. The facts of association whose contingencies one with another can be determined are the aspects of association which have here been enumerated.

Among the contingencies of associational phenomena one upon another which admit of statistical determination the following are especially significant.

The strength of association may vary with extent or with duration. It may be found that cohesion increases to a certain determinable point more rapidly than extent increases, and beyond that point less rapidly. For example, as a fact of observation, large states, large towns, large families are generally more coherent than small ones. A similar relation may be found between cohesion and duration.

Again, within limits, the stability of a group may be unaffected by a mobility of its units which permits individual units to disappear from the group and other units from without to replace them. Beyond a determinable limit such mobility of units may impair group stability.

The contingencies of associational phenomena with the degrees and modalities of composition are numerous and highly important.

Within determinable limits similar reactions of associated units are contingent upon other similarities of the units. Like units, in other words, tend to react in like ways. When sorts are combined in a mixture, the units of a sort may react in ways different from the ways in which they react when not in composition. And the effects of composition upon reaction may be a consequence in part of the proportions in which sorts are combined.

Inertia or momentum of associational reaction increases in a determinable ratio with the modality of variegation, that is, as one sort tends to dominate a composition. This is a familiar fact of our social life in all its phases, from fashion to politics. Transformation goes on at an increasing rate, which may be determined, as the proportion of variants from mode increases.

And at this point contingency of associational reaction upon external fact is discovered.

Adaption to environment or circumstance increases modality. Crisis multiplies variants.

Relations of toleration, the reactions of conflict and the reactions of adjustment are notoriously contingent upon forms of association, and these contingencies in a great number of instances admit of quantitative determination. The contingency of toleration is highest when association is tangentail. Conflict is most acute when associations are intersecting. Adjustments, both of the interests of units one with another and between the opposing tendencies towards modality and towards variability, are contingent upon the development of intra-secting association.